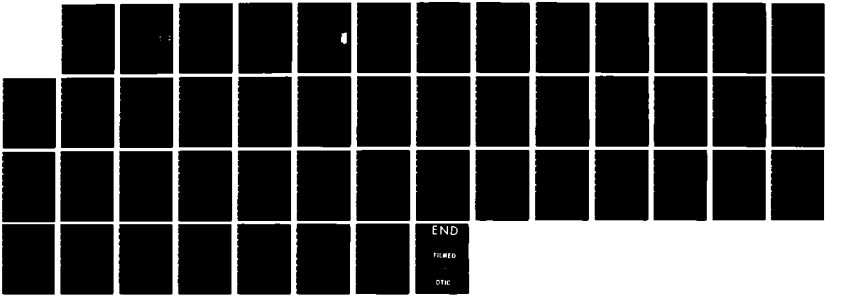
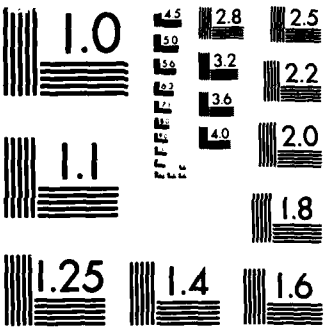
AD-A160 819 AN INVESTIGATION INTO PEER AND SUPERVISOR DIFFERENCES 1/1
IN THE OBSERVATION O (U) AIR FORCE INST OF TECH
WRIGHT-PATTERSON AFB OH M W DALEY DEC 85
UNCLASSIFIED AFIT/CI/NR-85-132T F/G 5/9 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

UNCLASS
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFIT/CI/NR 85-132T	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): An Investigation Into Peer And Supervisor Differences In The Observation Of Performance-Related Behaviors	5. TYPE OF REPORT & PERIOD COVERED: THESIS/DISSERTATION
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Mary W. Daley	8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS: AFIT STUDENT AT: Bowling Green State University	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: AFIT/NR, WPAFB OH 45433 - 6583	12. REPORT DATE: December 1985
	13. NUMBER OF PAGES: 39
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report): UNCLASS
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

DTIC ELECTE NOV 4 1985 S D B

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

APPROVED FOR PUBLIC RELEASE: IAW AFR 190-1

LYNN E. WOLAVER
Dean for Research and
Professional Development
AFIT, Wright-Patterson AFB

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

ATTACHED

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

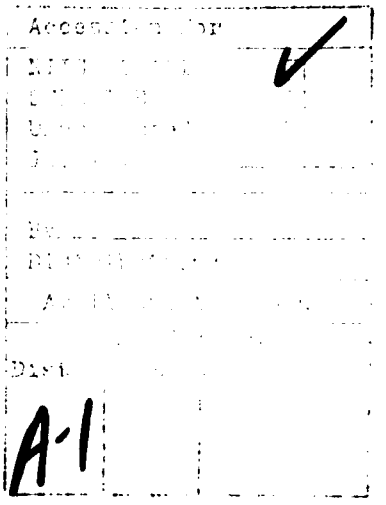
UNCLASS
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

DTIC FILE C AD-A160 819

85 11 04 069

ABSTRACT

The present study examined differences between peer and supervisors with respect to the observation of performance-related behaviors using the Instanteneous Report of Judgments (IRJ) technique. Undergraduate students and graduate teaching assistants served as peer and supervisor subjects respectively. All subjects viewed a videotape of an undergraduate giving a major classroom oral presentation. It was hypothesized that peer and supervisor subjects would differentially observe imbedded incidents due to their different role relationships to the presenter. The results indicated partial support for the hypothesis, with supervisor subjects observing significantly more critical incidents. An alternative interpretation for the results, that of rater expertise, is forwarded and discussed. Directions for future research in the area of cognitive processes and performance appraisal are discussed.



85-132T

AFIT RESEARCH ASSESSMENT

The purpose of this questionnaire is to ascertain the value and/or contribution of research accomplished by students or faculty of the Air Force Institute of Technology (AU). It would be greatly appreciated if you would complete the following questionnaire and return it to:

AFIT/NR
Wright-Patterson AFB OH 45433

RESEARCH TITLE: An Investigation Into Peer and Supervisor Differences in the Observation of Performance-Related Behaviors

AUTHOR: Mary W. Daley

RESEARCH ASSESSMENT QUESTIONS:

1. Did this research contribute to a current Air Force project?

( ) a. YES ( ) b. NO

2. Do you believe this research topic is significant enough that it would have been researched (or contracted) by your organization or another agency if AFIT had not?

( ) a. YES ( ) b. NO

3. The benefits of AFIT research can often be expressed by the equivalent value that your agency achieved/received by virtue of AFIT performing the research. Can you estimate what this research would have cost if it had been accomplished under contract or if it had been done in-house in terms of manpower and/or dollars?

( ) a. MAN-YEARS _____ ( ) b. $_____

4. Often it is not possible to attach equivalent dollar values to research, although the results of the research may, in fact, be important. Whether or not you were able to establish an equivalent value for this research (3. above), what is your estimate of its significance?

( ) a. HIGHLY SIGNIFICANT ( ) b. SIGNIFICANT ( ) c. SLIGHTLY SIGNIFICANT ( ) d. OF NO SIGNIFICANCE

5. AFIT welcomes any further comments you may have on the above questions, or any additional details concerning the current application, future potential, or other value of this research. Please use the bottom part of this questionnaire for your statement(s).

NAME GRADE POSITION

ORGANIZATION LOCATION

STATEMENT(s):

FOLD DOWN ON OUTSIDE - SEAL WITH TAPE

AFIT/NR
WRIGHT-PATTERSON AFB OH 45433

OFFICIAL BUSINESS
PENALTY FOR PRIVATE USE, $300

NO POSTAGE
NECESSARY
IF MAILED
IN THE
UNITED STATES

BUSINESS REPLY MAIL
FIRST CLASS PERMIT NO. 73236 WASHINGTON D.C.

POSTAGE WILL BE PAID BY ADDRESSEE

AFIT/ DAA
Wright-Patterson AFB OH 45433

FOLD IN

AN INVESTIGATION INTO PEER AND SUPERVISOR DIFFERENCES IN THE OBSERVATION OF PERFORMANCE-RELATED BEHAVIORS

Mary W. Daley

Submitted to the Graduate College of Bowling Green State University in partial fulfillment of the requirements for the degree of

MASTER OF ARTS

December 1985

Approved:

Advisor

ACKNOWLEDGEMENTS

The author wishes to acknowledge all those persons who lent their support, guidance, and efforts toward the successful completion of this thesis. Many thanks go to my chairman, Dr. William Balzer, for his constant guidance and support throughout all stages of this project. Special thanks are also extended to my other committee members, Dr. John Cavanaugh and Dr. Charles Cranny for their helpful suggestions and criticisms. A big thanks must go to many of my graduate student collegues who served as subjects in this experiment. And finally, I wish to express a heartfelt thanks to some special people -- Patti Parkison, Lorne Sulsky, Nancy Grevengoed, and Wade Gibson -- for their ubiquitous emotional and intellectual support.

TABLE OF CONTENTS

LIST OF TABLES

Performance appraisal has been a major area of research in industrial psychology for a number of years (cf., Landy & Farr, 1980). Appraising an employee's performance is essential to an organization for a number of reasons, including organizational staffing, distributing rewards and sanctions, and evaluating motivational and educational programs (Landy & Farr, 1983). In spite of the amount of time and money spent on implementing and improving performance appraisal systems, however, performance appraisals are continually plagued with error. Unsolved problems of rater biases, poor rating scales, and lack of understanding of the behavioral constructs being measured all call for continued research in performance appraisal (Kane & Lawler, 1978).

In an effort to understand and minimize rating system errors, researchers have begun viewing performance appraisal as a special case of the general person perception process, where performance-related information about the ratee must be gathered and integrated by a single (or multiple) rater(s) (Borman, 1978; Feldman, 1981; Landy & Farr, 1983; Wexley & Klimoski, 1984). Some empirical work has begun to use the person perception process as the framework for understanding the performance appraisal process (Balzer, in press; Banks, 1979; Nathan & Lord, 1983). To date, however, there has been no published empirical research examining how the position of the rater to the ratee (e.g., supervisor, peer, subordinate,

client, or self) influences the person perception process during performance appraisal (Landy & Farr, 1980, 1983). This study applies the person perception process to rater-ratee relationships in appraisal situations by examining differences in how supervisors and peers observe performance-related information.

Prior to describing the research study, a brief review of the literature on performance appraisal will be presented, followed by a discussion of research investigating rater-ratee relationships. Next, literature focusing on the cognitive processes involved in performance appraisal will be discussed with special attention given to the earlier stage of observation, which this study will investigate. Following this, techniques for measuring observation will be reviewed and evaluated. Finally, an overview of the present study will be presented and hypotheses stated.

Performance appraisal

Every organization finds it necessary to periodically evaluate their employees against certain standards or criteria of satisfactory performance. Thus, the study of performance appraisal is necessary to aid organizational and employee effectiveness and productivity. The appraisal context, which refers to the purpose or reason for doing the evaluation, may be the most pervasive influence on the performance appraisal. DeNisi, Cafferty, and Meglino (1984) and Bernardin (1978) cite considerable evidence that ratings

accomplished for administrative purposes (e.g. promotion, selection, etc.) are more lenient than those provided for research purposes. However, the majority of studies have been laboratory investigations (Landy & Farr, 1980); studies must be conducted in other contexts before even tentative conclusions can be drawn.

Various rating formats have been shown to differently reflect workers' true performance. For example, trait rating scales have been shown to be prone to errors due to the global nature of the categories (Borman & Dunnette, 1975). More specific scales such as behaviorally-anchored rating scales (BARS, Smith & Kendall, 1963) are improvements over trait scales, but do not eliminate all problems (Borman & Dunnette, 1975). Research has also shown that training raters has some effect on ratings. Most research in this vein focuses on whether training reduces various rater errors. Some studies have found that the number of rating errors substantially decreases following training (Bernardin & Walter, 1977; Bernardin, 1978; Brown, 1968). Of practical concern, however, is how long training effects last. Latham, Wexley, and Pursell (1975) found a reduction in rating errors up to six months after training; however, Bernardin (1978) and Warmke and Billings (1979) reported no long-term effects for training.

Research has also looked at rater and ratee characteristics and their interactions. Included here is

research on rater consistency across and within task (Borman, 1977, 1979), race and sex effects (Norton, Guslafson, & Foster, 1977), rater knowledge of ratee and job (Borman & Dunnette, 1975), and rater-ratee role relationship (Borman, 1978). Generally, these studies provide few definitive answers as to how rater and ratee characteristics influence ratings or offer few conclusions for improving ratings. Research in the rater-ratee role relationship area, however, may shed some light on how the varied perspectives that different raters possess may be used to improve ratings.

Rater-ratee role relationship. A variety of research has reported that who performs the performance appraisal does make a difference. There are many possible relationships between the rater and the ratee. Ratings may be accomplished by an individual's supervisor, peers, subordinates, client or even by the individual him or herself. Landy and Farr (1980) conclude that the different rater roles (peer, supervisor, subordinate, self, etc.) may each have unique vantage points of a given ratee's performance, and that each view may contribute differently in assessing performance. The rater's role may lead to possible differences in motivation, knowledge, observational opportunities, or purposes on the part of the rater.

Supervisor's judgments comprise the vast majority of ratings completed within organizations (Campbell, Dunnette, Lawler, & Weick, 1970). Whitla and Tirrell (1953), Mandell

(1956), and Zedeck and Baker (1972) investigated supervisor ratings at various organizational levels with respect to leniency and validity. Mandell (1956) found no significant difference between immediate and higher level of supervisors of workers with respect to leniency. Zedeck and Baker (1972) and Whitla and Tirrell (1953) both found convergent validity for various levels of supervisor ratings, the former with nursing personnel and the latter with flight mechanics.

Peer assessment, whether via nominations, rankings, or ratings have also been the focus of much research, but have yet to gain widespread use (Kane & Lawler, 1978). Reliability of peer ratings has been demonstrated by consistent ratings over time and across work groups (e.g., Gordon & Medlund, 1965; Wherry & Fryer, 1949). In an empirical study using police officers and sergeants, Love (1981) found that all three methods of peer assessment had significant reliability and validity. Furthermore, validity coefficients obtained were not significantly biased by friendship between peer assessors. Validity of peer ratings against criteria of subsequent promotion and performance ratings have also been demonstrated. Roadman (1964) obtained peer ratings on 13 work and personal characteristics of 56 managers attending a month-long business training session. Two years later, 10 of the 13 characteristics were found to correlate significantly with the number of subsequent promotions. Kraut (1975) found that peer ratings of 83

executives obtained during a training session on Roadman's (1964) 13 characteristics were predictive of the executives' later performance appraisals.

Klimoski and London (1974) conducted a study that examined supervisor, peer, and self ratings. Subjects were 153 registered nurses who rated and ranked each other and themselves on a given scale. In addition, nursing supervisors also rated and ranked these nurses. The researchers concluded, based upon factor analytic results, that each source did rate performance from a distinctly different perspective. Specifically, peer evaluations were more likely to focus on task relevant abilities and competencies in appraising others, while supervisors may have considered additional information to get a broader perspective.

Limited empirical research has been conducted on raters in subordinate role relationships to the ratees. Research evidence is equivocal, showing both significant agreement between superior and subordinate ratings (Graen, Dansereau, & Minami, 1972) and no relationship between ratings of foremen and workers (Besco & Lawshe, 1959). Teacher evaluation studies (e.g., Nathan and Lord, 1983) have used subordinates (students) as raters. However, subordinate raters are not used in these studies as means to investigate the rater-ratee role relationship but rather as incidental samples.

There has been limited theoretical research into rater-

ratee relationships. Mumford (1983) forwarded Social Comparison Theory to explain differences in evaluations related to the rater's role relationship to the ratee. Social Comparison Theory holds that

> ... individuals are motivated to evaluate their opinions and abilities, and when objective information concerning the adequacy of their opinions and abilities is not available they will attempt to obtain such information by comparing their opinions and abilities to those held by other individuals .... individuals who are similar to the person making the comparison will be preferred as standards for evaluation .... (Mumford, 1983, p. 874)

If individuals are competing under conditions where objective performance evaluations occur infrequently and where rewards are seen as strongly contingent upon task relevant abilities and competencies, evaluation by comparison with similar others (peers) is likely. Observation by peers, then, is likely to be far from random, focusing on specific abilities and competencies while a supervisor rater who is not operating within the same social level as the ratee will observe incidents pertinent to his or her role as a supervisor.

Borman (1978) notes that a rater's role may place constraints upon his or her opportunity to view relevant ratee behaviors. For example, supervisors may view certain

behaviors of the subordinates but peers are not subject to the same constraints and thus may see different behaviors. The rater's role may also influence the motivation behind the rating. For example, a peer's motivation may be simply to follow the company's policy or it may be to get even with a coworker or even to help out a friend. The supervisor's motivation may be to follow company policy, help a good subordinate along to a promotion, or to simply accomplish the rating task. Finally, prior commitment to a subordinate by a superior may influence the rating process. Bazerman, Beekman, and Schoorman (1982) found that if an initial commitment had been made by the rater to the ratee, the rater will subsequently bias his or her ratings to justify that support.

Cognitive processing model of performance appraisal

A number of researchers in the area of performance appraisal have begun to develop cognitive models to describe the process by which information is heeded and used to make performance appraisal decisions (Borman, 1978; DeNisi, Cafferty, & Meglino, 1984; Feldman, 1981). For example, the DeNisi et al. (1984) model proposes that performance appraisal is accomplished via a set of social cognitive operations. Generically, these steps are observation, coding, storage, retrieval, and integration. Observation is defined as a conscious or nonconscious attending to specific performance rating information. The next step, that of

coding, may be defined as the transformation of "raw" stimulus information into mental representations. Storage involves the retention of mental representations relevant to a later task, such as remembering employee performance until it is time to conduct an appraisal. Next, retrieval is the calling up of relevant information from long-term memory. The final stage, that of integration, consists of weighting and combining all relevant information from memory to make a useful and valid rating judgment.

Error may be introduced at any stage in human information processing (DeNisi et al., 1984). Each stage has limitations because humans can only process a small sample of the variety of stimuli available to them. Errors may first be introduced at observation; influences such as preconceived notions of the ratee, purpose of the rating, rating instrument format, and time pressure at rating may enter here and affect what the rater observes (DeNisi et al., 1984). Later stages are affected by errors, compounding error introduced at an earlier stage. For example, Balzer (in press) showed that one's expectations about a ratee may influence how performance-related information is encoded, a finding consistent with research in the area of cognitive schemas (Hastie, 1981; Fiske & Taylor, 1984; Taylor & Crocker, 1981). Storage may be affected by the rater's schemas or categories in that information consistent with a schema is likely to be stored in memory while inconsistent

and irrelevant information may be given less weight or possibly not committed to memory (Taylor & Crocker, 1981). DeNisi et al. (1984) also suggest that attributions of ratee performance, implicit theories about a ratee, and the salience of performance information may induce error at storage and at later stages.

Retrieval, too, can be affected by schemas. Here, general characteristics of the schema associated with the ratee at storage may be used to retrieve information, resulting in global evaluations consistent with the schema (DeNisi et al., 1984; Feldman, 1981). Finally, inaccurate weighting and combining of stimulus information may cause error at the integration stage (Hamilton & Huffman, 1971).

In summary, the rater's rating process may be thought of as a set of interrelated cognitive tasks. Biases are present in each of the information processing stages, and biases in earlier stages are likely to be compounded in later stages due to the dependence on initially biased information. Thus, a better understanding of what happens during the first stage in the cognitive process has definite implications for the use of information in the later cognitive stages. The present research investigated the way different types of raters observe performance rating information.

Observation. In order to appraise another individual's performance, the rater must first observe job-related behavior or behavior products such as reports or material

produced. These are the necessary initial stimuli to the rater's performance-rating process. However, the rater is unable to use all information but must sample from available behaviors (March & March, 1978), and what is sampled in this observation stage has been shown to affect rating accuracy (Murphy, Garcia, Kerkar, Martin, & Balzer, 1982).

Fiske and Taylor (1984) reviewed the role of observation as theorized and researched in the area of social cognition. A number of factors are thought to affect the information which a perceiver (or rater) is expected to observe, including the salience of the stimulus relative to its context (e.g., a female in a previously all-male work group), the vividness of the stimulus itself, independent of context, and the environmental cues (appraisal context, mentioned earlier). For example, a supervisor who has been trained to recognize how stereotypes may influence ratings may not view one female in a previously all-male work group as salient while the members of the work group may view her every action as salient and attribute much to her uniqueness. Thus, what the supervisor and peers observe about the same person may be quite different.

DeNisi et al. (1984) review much of the social cognitive work in this area and conclude that the raters are active seekers of information at this stage. What the raters observe (and ultimately encode, store, retrieve, and integrate) is determined by preconceived notions about the

ratee, appraisal purpose, nature of the rating scale, and time pressures. Preconceived notions or impressions by the rater of the ratee may affect what is observed by activating schema and causing a search for information to test that schema's validity. DeNisi et al. (1984) cite many studies that suggest that the administrative purpose (promotion, selection, etc.) for the performance appraisal results in more lenient evaluations than those conducted for research purposes. Knowing the purpose of the rating beforehand may allow the rater to judge how thorough he or she must be when observing a rater's performance. The nature of the rating instrument may serve to guide the rater to look for certain dimensions and not others. Finally, time pressures may influence what is observed, causing the rater in time pressured rating situations (e.g., when there are many conflicting demands upon the rater's time) to seek only negative information (DeNisi et al, 1984).

In addition to the social cognitive work on observation, some work has been done in the performance rating area. Landy and Farr (1980) and others (Kane & Lawler, 1978; Lewin & Zwany, 1976) intimate that the difference between peer and supervisory ratings may be partly due to the different ways that peers and supervisors process information (e.g., differences in the salience or vividness of a stimulus to a peer or supervisor). Thus, research investigating the cognitive processes involved in making

performance appraisals may directly contribute to explaining rating differences.

In summary, the literature in several areas suggests that cognitive processing differences as well as external demands may be contributing to the differences in ratings provided by individuals participating in different roles. Therefore, it seems worthwhile to investigate whether rater-ratee relationships do affect observation during performance evaluation. In this study, the rater-ratee relationship was manipulated by using peers and supervisors in the rater role. Prior to investigating observational differences in performance ratings, various strategies used to measure observation will be discussed.

Measurement of observation

A number of different strategies have been used to measure observation in psychological research, with some being more practical than others for measuring observation in a performance appraisal situation. Physiological measures, such as tachistoscopic or dichotic presentation of stimuli, pupil dilation, and time spent viewing discrete stimuli have been used in experimental psychology to measure observation (Klatzky, 1980). However, these measures are difficult to collect in a performance rating environment, and it is still uncertain whether these are measures of perception on the part of the subject or due simply to physiological arousal (Banks, 1979).

More natural, but indirect, measurement techniques have also been used to measure observation. Diary-keeping, where raters record crucial behavioral events as they occur, is one method used to measure rater's observation (Bernardin & Walter, 1977). However, even though raters may observe similar behaviors, biases such as preconceived notions may still affect what behavioral incidents are entered into the diary (Balzer, in press). Recall tasks may help researchers understand the observational process that a rater is using by simply asking the rater to recall specific behavioral incidents which he or she remembers (Fiske & Taylor, 1984). This method, however, moreso than diary-keeping, is subject to distortion due to the time delay between observation and measurement (Ericsson & Simon, 1980). In addition, both diary-keeping and recall may require the use of later cognitive stages (e.g. coding, storage, and retrieval) and thus may be impure measures of observation.

Policy capturing is yet another measure used to determine aspects of an individual's cognitive processes (Hoffman, 1960). This measure uses regression analysis to investigate how a rater is using information when making a judgment (Zedeck & Kafry, 1977). Though sometimes suggested as a mesaure of observation, policy capturing may more appropriately be a measure of integration.

A new and promising measurement technique, the Instantaneous Report of Judgments (IRJ), is purported to

capture the ongoing process of observation without disturbing it (Banks, 1979). Using this technique, raters view a videotape of a target ratee performing some task and are asked to attend to a specific area of performance. Whenever a rater observes performance behaviors related to this prespecified area of performance, the rater presses one of several buttons which both "marks" the location on the tape where the observation is made while simultaneously indicating the rated level of effectiveness of the specific behavior. Raters are also asked to verbalize their observations to help indicate further which piece of performance-related information was being observed.

The IRJ technique has several advantages over the previously mentioned techniques for measuring observational processes in a performance appraisal context. First, the rater identifies any information he or she considers relevant for evaluating performance, unlike some other methods where preconceived cues must be used. Secondly, by reporting observation and judgments as they occur, raters have full control over the task, which minimizes demand characteristics and experimenter bias. Finally, since the IRJ technique allows for instantaneous reporting, there is less opportunity for the later stages of encoding, retrieval, and integration to be confounded in this indirect measure of observation.

Measuring an internal cognitive process such as observation is difficult, however, since any

operationalization of observation is only an indirect measure. The IRJ technique rests on the belief that button-pressing can measure the ongoing cognitive process of observation and encoding without disturbing or changing that process. Newtson (1976) argues that ongoing behavior is segmented into discrete chunks of events during perception, with segmentation occurring at natural "breakpoints" or changes in the action. Subjects mark such breakpoints in a stream of onging behavior while performing the IRJ task. Thus, button-pressing may actually reflect when performance-related observations in ongoing behavior occur. Ebbesen (1980) takes issue with Newtson's idea that button-pressing taps a basic observation/encoding process, however. He posits that button-pressing may reflect a secondary process added to the "normal" cognitive process. Information may be continuously encoded rather than in "chunks" as Newtson theorizes. Additionally, the requirement to button-press itself may be a demand characteristic that interferes with the continuous coding of information.

Thus, although there may be some limitations in using the IRJ method, it remains a promising technique for measuring observation. By requiring subjects to respond as the behaviors occur, the technique minimizes confounding due to later stages of information processing. For this reason, it was chosen as the measurement method for this study.

Design and Hypothesis

This study focused on differences between peers and supervisors in observing performance-related behaviors. Specifically, subjects in peer and supervisory role relationships to an undergraduate student viewed a videotape featuring the student giving a class presentation where specific critical behaviors were imbedded in the tape. Subjects in both peer and supervisory relationship to the student were instructed to attend to the same targeted behavioral dimension. Subjects were asked to respond using a slight variation of the IRJ procedure. Post-experimental analyses were conducted to determine whether peers and supervisors observed different behavioral incidents.

Based on literature reviewed in the introduction, it was hypothesized that given the same standardized behaviors, peers and supervisors would differently observe specific behavioral incidents. This was predicted because peers and supervisors may process information differently and hence heed different input information (Kane & Lawler, 1978; Landy & Farr, 1980; Lewin & Zwany, 1976).

Method

Subjects

Thirty undergraduate students and thirty graduate student teaching assistants (TAs) served as raters for this study. Undergraduate students served in the role of peers while the TAs served in the role of supervisors. Undergraduate students received experimental credit for their

participation; graduate student participation was voluntary. Thirty of the subjects were male and thirty were female. In the peer condition, seventeen subjects were male and thirteen were female. Among the supervisor subjects, thirteen were male and seventeen were female.

The particular number of subjects assigned to each group (peer or supervisor) was determined by practical considerations of subject availability. Theoretical concerns were also considered in this decision. For a significance level of **a** = .05, an estimated medium effect size, and an **n** (per group) = 30, the statistical power of the between group comparison *t*-test is 0.61 and for the chi-square test is 0.64. To attain the generally recommended power of .80 at **a** = .05, 100 subjects (50 in each group) for the *t*-test and 87 total subjects for the chi-square test would be needed (Cohen, 1977). Such a large sample size is impractical due to a limited number of graduate teaching assistants.

Stimulus Materials

One 10 minute videotape of an undergraduate student giving a classroom presentation in an introductory psychology course was prepared for use in this experiment. The "student", portrayed by a female actor, gave an oral presentation on the effect of subjective perceptions of price on consumer behavior. The videotaped lecture was constructed to contain critical behavior incidents adopted from the Nathan and Lord (1983) stimuli, based upon the five

performance dimensions developed by Harari and Zedeck (1973) for evaluating performance: Organization, Delivery, Relevance, Interpersonal Relations with Students, and Depth of Knowledge. Six of these incidents were related to the target dimension of Depth of Knowledge. Both positive and negative incidents in all dimensions were represented. The videotape script was read by five industrial-organizational psychology graduate students prior to filming. These students were given a list of the five dimensions and asked to categorize all scripted behavioral incidents into one of the five dimensions. The graduate students also viewed the videotape and categorized observed behavioral incidents into one of the five dimensions. This was done to allow for dynamics of the actual acting out of the behaviors to be taken into account, if important. The results showed that the confederates (I/O graduate students) did categorize the behavioral incidents into the proper dimensions.

Procedure

Both undergraduate (peer raters) and graduate (supervisor raters) students participated individually in the present study. Prior to viewing the videotape, each subject was given instructions and brief (5-10 minutes) training in differentiating between behavioral incidents and evaluations, as well as definitions and examples of the five performance dimensions (see Appendices A and B). Peer rater subjects were told that they would be viewing a videotape in which a

fellow undergraduate student is giving a required presentation on a topic in introductory psychology. Supervisor rater subjects were told that they would be viewing a videotape of an undergraduate student giving a required classroom presentation and that they were to assume the role of teaching assistant (supervisor) as they viewed the tape. All subjects were instructed to attend to the targeted dimension of Depth of Knowledge. Subjects were told that they will be asked to rate the presenter on all dimensions after the experiment. Subjects were instructed to stop the videotape by pushing the button on a control panel in front of them anytime they observed a behavioral incident relating to the targeted dimension. Subjects were also instructed to simultaneously verbalize the particular incident (which was tape recorded) and to rate the incident using a four-point scale (1 = very negative, 4 = very positive). The experimental apparatus (stop/start tape mechanism, tape recorder) was explained and demonstrated to each subject. After answering any questions, the experimenter started the videotape. At the completion of the videotape lecture, the subject was asked several questions by the experimenter (Appendix C). This question and answer session was also tape recorded. After completion of the experimental task, the subject was thanked for his or her participation and debriefed (see Appendix D).

Analysis

The hypothesis was analyzed via the chi-square and *t*-test statistics (Hays, 1981). The dependent variable of observance of behavioral incidents was recorded by noting the number displayed on the videocassette recorder (VCR) counter when the subject stopped the tape to record his or her observation. Post-experimental analyses of the videotape and tape recording resulted in coding of observance of behavioral incidents as "1" and nonobservance as "0". Each Depth of Knowledge behavioral incident from the videotaped lecture was analyzed with the chi-square statistic to see if observance differed between peer and supervisor subjects, with the count of observations or nonobservations in appropriate cells. As an alternative analysis, a *t*-test was performed to test differences in the number of observations of behavioral incidents between peers and supervisors across collapsed behaviors within the Depth of Knowledge dimension.

Results

The hypothesis that peers and supervisors would differently observe specific behavioral incidents was analyzed via the chi-square statistic for the imbedded Depth of Knowledge critical incidents. Six separate chi-square analyses were done, one for each critical incident, and are reported in Table 1. An examination of Table 1 shows that significant chi-square values were found for three of the six critical incidents. Differences between groups was greatest

Table 1

Observation Count, Chi-Square, and Probability Values for each Depth of Knowledge Critical Incident

Behavior	PNO	PO	SNO	SO	$\chi^2(1)$	p
BEH1 - Presenter does not know the name of the writer critical of universality of the Law of Demand. (Negative)	8	22	5	25	0.39	.531
BEH2 - Presenter responds with multiple articles when asked for information on the effect of price on purchasing behavior. (Positive)	2	28	0	30	0.52	.472
BEH3 - Presenter is very familiar with research on student perceptions of quality as a function of price only. (Positive)	26	4	16	14	6.43	.011
BEH4 - Presenter is able to elaborate differences in experimental results for male and female product preference as a function of product price. (Positive)	11	19	3	27	4.57	.033
BEH5 - Presenter does not know how Jacoby determined the differences in the quality of beers used in his multicue study. (Negative)	2	28	0	30	0.52	.472
BEH6 - Presenter is not familiar with the topic of Absolute Price Threshold. (Negative)	18	12	4	26	12.13	.001

Note: N = 60
PNO - Peer Not Observed
PO - Peer Observed
SNO - Supervisor Not Observed
SO - Supervisor Observed

with respect to the sixth behavioral incident in the videotape. Referring to Table 1, supervisor subjects observed Behavior 6 more than twice as often as did peer subjects, yielding a Yates' corrected chi-square value of 12.13 (1, N = 60), p = .033. A significant corrected chi-square value of 6.43 (1, N = 60), p = .011 was obtained for Behavior 3. Here, fourteen supervisor subjects observed the incident while only four peer subjects observed the same incident. Finally, a significant corrected chi-square value of 4.57 (1, N = 60), p = .033 was obtained for the fourth incident presented in the videotape. Again, supervisor subjects more often observed the behavioral incident than did peer subjects. An alternative but identical test of the hypothesis was to compute a t-test between supervisor and peer conditions examining the mean number of behavioral incidents observed. Collapsing across the six behaviors within the Depth of Knowledge dimension, the t-test reported in Table 2 showed that supervisors observed significantly more incidents than did the presenter's peers (Ms = 5.10 and 3.77, respectively; t (58) = 5.36, p < .001).

In summary, drawing from the results of the six separate chi-square tests and the t-test, partial support was found for the hypothesis that observational differences exist between peers and supervisors on given behavioral incidents. That is, for the particular incidents in this experimental

Table 2

Peer and Supervisor Differences in Observation on all Depth of Knowledge Incidents

Condition	N	M	SD	t-Value	DF	2-tail Prob.
Peer	30	3.77	0.94	5.36	58	.000
Supervisor	30	5.10	0.94			

condition, supervisors were more observant than peers of all imbedded incidents.

Though data were collected on the ratings given on the individual Depth of Knowledge behaviors observed and on the overall ratings given the presenter on each of the five dimensions, these analyses are not presented here. These analyses, along with the subjects' responses to the post-experimental narrative questions (Appendix C) are available upon request from the author.

Discussion

The hypothesis that peers and supervisors would differently observe specific behavioral incidents was partially supported. Results from the chi-square analyses showed that three of the six behavioral incidents imbedded on the videotape were observed significantly more often by graduate students role playing teaching assistant-supervisors than undergraduates role playing classroom peers. The six imbedded critical behavioral incidents were, necessarily, constructed to be more distinct than many other actions in the videotape. Such distinctiveness or salience may be a major condition for attracting a perceiver's attention. The behavioral incidents that were observed significantly more often by the teaching assistants/supervisors than by the peers may be less salient than the other incidents. Differential salience of stimuli dependent upon the rater's role relationship to the ratee has been suggested by Kane and

Lawler (1978) and Lewin and Zwany (1976). However, no data are available to test whether some critical behaviors are indeed more salient than others.

However, supervisors did not observe all incidents within the Depth of Knowledge dimension significantly more than peers (see Table 2). The nature of the stimulus material and the manipulation of role may help to explain this pattern of findings. The stimulus material consisted of the presentation of academic material, that is, teaching behavior. Although the presenter in the videotape was an undergraduate and the fact that instructions to both peer and supervisor subjects stressed that this was an undergraduate class presentation, some of the peers and supervisors may have perceived the presenter as a teacher, despite the intended role manipulation. Some evidence for this is available from responses to the narrative question which asked whether the subject believed his or her role relationship to the presenter (peer or supervisor) influenced what he or she observed. Approximately thirty-five percent of the subjects responded that the presentation was more characteristic of teacher rather than student performance. One third of the undergraduates (peer condition) reported that they were observing someone as a teacher (supervisor) rather than as a fellow undergraduate. Approximately eighty-seven percent of the graduate student teaching assistants (TAs) were able to see the undergraduate in the videotape as

a subordinate, even though there was a tendency by approximately thirteen percent of the TAs to judge the undergraduate against tougher instructor or peer (TA) standards (e.g., more in-depth discussion of cited articles). Thus, the stimuli used in this study may have weakened the intended role manipulation, especially for subjects in the peer condition.

Although the results may be interpreted as partial support for the role difference hypothesis, different levels of rater expertise may be an important and viable alternative explanation. For the three significant Depth of Knowledge incidents, all showed greater observation by supervisors than by peers. In addtion, for the three nonsignificant Depth of Knowledge behaviors, supervisors observed the incidents more often than did the peer subjects. In all instances, then, the supervisors, who had more expertise in teaching than peers, were more observant of the imbedded behavioral incidents.

The importance of the rater's expertise or job knowledge has received some attention in the performance appraisal literature (Amir, Kovarsky, & Sharan, 1970; Whitla & Tirrell, 1953; Zedeck & Baker, 1972). These studies have found expertise to be related to prediction of subordinates' future performance. They do not, however, integrate the expertise issue into the information processing model of performance appraisal. The findings in this study suggest

that rater expertise may be an important factor in determining what raters observe. All the supervisors had had experience teaching and thus were familiar with the task-relevant abilities that the undergraduate presenter in the videotape should possess. Thus, the supervisors may have had more expertise observing teaching behaviors than the peer subjects and thus were perhaps more attentive observers of the undergraduate's teaching behaviors (Hastie, 1982; Fiske & Taylor, 1984).

While the data are supportive of the rater expertise explanation, the confounding of expertise with role relationship makes it impossible to draw any firm conclusion. Thus, it is not clear whether the difference in observation is due to the role relationship of the rater to the ratee or to his or her expertise in the rated area. The finding of observational differences between peers and supervisors may be consistent with Landy and Farr's (1980) contention that persons in different rater roles may have unique perspectives based upon their role and hence contribute different information to performance assessment. But the extent to which that differential role perspective is due to rater expertise is unknown.

The results of this study must be interpreted with caution. First of all, operationalization of an internal process such as observation is a difficult task. This research only measured outcomes (button pressing behavior)

and inferred observational processes from these responses. Although the Instantaneous Report of Judgments (IRJ) measurement technique allows for instantaneous reporting, it remains only an indirect measure of a complex process.

There are also a number of methodological limitations in the present study. The experimenter's presence during the experiment may have contributed to unknown demand characteristics such as the subject paying more attention to the behaviors than would be possible in a more realistic work setting. The role manipulation may have been weakened by the stimulus itself being quite characteristic of teacher behavior. The stimulus material (videotape) was developed by researcher with some input from graduate students. Although most of the development was accomplished with undergraduates, the extent to which researchers' and graduate students' knowledge of teaching behaviors confound the findings is not known. Finally, the use of undergraduate and graduate student subjects, though appropriate for the task of observing and evaluating teaching behavior, does limit the generalizability of these findings. Future research in this area should examine more real world work behaviors, using salespersons, business executives, etc. as subjects.

Implications of this study are of both theoretical and practical significance. Theoretically, this study adds to the limited performance appraisal literature investigating the cognitive process of observation by offering the variable

of expertise as a possible factor affecting observation. Practically, the finding that more experienced personnel (i.e., supervisors who are very familiar with the subordinate's tasks) may observe more of the nuances of job-related behavior may be used to design performance appraisal systems and which can take advantage of both the different information observed by persons in various role relationships to the ratee as well as rater expertise.

The results of this study stimulate suggestions for future research. Investigation into the role that expertise plays in observation and later cognitive processing stages is needed. Studies which simultaneously manipulate both expertise and rater role relationship may help clear up some of the present confusion concerning the relationship between these two variables. Further research using the Instantaneous Report of Judgments (IRJ) technique is also suggested. Its use can be fairly simple and it may indeed tap into information not available via other methods. Research comparing the IRJ technique with other methods may discover whether any information is uniquely measured by various methods.

References

Amir, Y., Kovarsky, Y., & Sharan, S. (1970). Peer nominations as a predictor of multistage promotions in a ramified organization. Journal of Applied Psychology, 54, 462-469.

Balzer, W.K. (in press). Biased recording of performance-related behaviors: The effects of initial impression and centrality of the appraisal task. Organizational Behavior and Human Decision Processes.

Banks, C.G. (1979). A laboratory study of the decision-making processes in performance evaluation. Unpublished PhD dissertation, University of Minnesota.

Bazerman, M.H., Beekman, R.I., & Schoorman, F.D. (1982). Performance evaluation in a dynamic context: A laboratory study on the impact of prior committment to the ratee. Journal of Applied Psychology, 67, 873-876.

Bernardin, H.J. (1978). Effects of rater training on leniency and halo errors in student ratings of instructors. Journal of Applied Psychology, 63, 301-308.

Bernardin, H.J., & Walter, C.S. (1977). Effects of rater training and diary-keeping on psychometric error in ratings. Journal of Applied Psychology, 62, 64-69.

Besco, R.O., & Lawshe, C.H. (1959). Foreman leadership as perceived by superiors and subordinates. Personnel Psychology, 12, 573-582.

Borman, W.C. (1979). Individual differences correlates of accuracy in evaluating others performance effectiveness. Applied Psychological Measurement, 3, 103-115.

Borman, W.C. (1978). Exploring the upper limits of reliability and validity in job performance ratings. Journal of Applied Psychology, 63, 135-144.

Borman, W.C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. Organizational Behavior and Human Performance, 20, 238-252.

Borman, W.C., & Dunnette, M.D. (1975). Behavior-based versus trait-oriented performance ratings: An empirical study. Journal of Applied Psychology, 60, 561-565.

Brown, E.M. (1968). Influence of training, method, and relationship on the halo effect. Journal of Applied Psychology, 52, 195-199.

Campbell, J.P., Dunnette, M.D., Lawler, E.E., III, & Weick, K.E., Jr. (1970). Managerial behavior, performance, and effectiveness. New York: McGraw-Hill.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences. New York: Academic Press.

DeNisi, A.S., Cafferty, T.P., & Meglino, B.M. (1984). A cognitive view of the performance appraisal process: A model and research propositions. Organizational Behavior and Human Performance, 33, 360-396.

Ebbesen, E.B. (1980). Cognitive processes in understanding ongoing behavior. In R. Hastie, T.M. Ostrom, E.B. Ebbesen, R.S. Wyer, Jr., D.L. Hamilton, & D.E. Carlston (Eds.), Person memory: The cognitive basis of social perception. Hillsdale, NJ: Erlbaum.

Ericsson, K.A., & Simon, H.A. (1978). Retrospective verbal reports as data. C.I.P. Working paper. No. 397, Carnegie - Mellon University.

Feldman, J.M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. Journal of Applied Psychology, 66, 127-148.

Fiske, S.T., & Taylor, S.E. (1984). Social cognition. Reading, MA: Addison-Wesley.

Gordon, L.V., & Medlund, F.F. (1965). The cross-group stability of peer ratings of leadership potential. Personnel Psychology, 18, 173-177.

Graen, G., Dansereau, F., & Minami, T. (1972). An empirical test of the man-in-the-middle hypothesis among executives in a hierarchical organization employing a unit-set analysis. Organizational Behavior and Human Performance, 8, 262-285.

Hamilton, D.L., & Huffman, L.J. (1971). Generality of impression formation for evaluative and non-evaluative judgments. Journal of Personality and Social Psychology, 20, 200-207.

Harari, O., & Zedeck, S. (1973). Development of behaviorally anchored scales for the evaluation of faculty teaching. Journal of Applied Psychology, 58, 261-265.

Hastie, R. (1981). Schematic principles in human memory. In E. Higgins, C. Herman, & M. Zanna (Eds.), Social cognition: The Ontario symposium (Vol. 1). Hillsdale, NJ: Erlbaum.

Hays, W.L. (1981). Statistics. (3rd ed.). New York: Holt, Rinehart and Winston.

Hoffman, P.J. (1960). The paramorphic representation of clinical judgment. Psychological Bulletin, 57, 116-131.

Kane, J.S., & Lawler, E.E. (1978). Methods of peer assessment. Pychological Bulletin, 85, 555-586.

Klatzky, P. (1980). Human memory: Structure and processes. San Francisco: W.H. Freeman and Company.

Klimoski, R.J., & London, M. (1974). Role of the rater in performance appraisal. Journal of Applied Psychology, 59, 445-451.

Kraut, A.I. (1975). Prediction of managerial success by peer and training-staff ratings. Journal of Applied Psychology, 60, 14-19.

Landy, F.J., & Farr, J.L. (1980). Performance rating. Psychological Bulletin, 87, 72-107.

Landy, F.J., & Farr, J.L. (1983). The measurement of work performance. New York: Academic Press.

Latham, G.P., Wexley, K.N., & Pursell, E.D. (1975). Training managers to minimize rating errors in the observation of behavior. Journal of Applied Psychology, 60, 550-555.

Lewin, A.Y., & Zwany, A. (1976). Peer nominations: A model literature critique and a paradigm for research. Personnel Psychology, 29, 423-447.

Love, K.G. (1981). Comparison of peer assessment methods: Reliability, validity, friendship bias, and user reaction. Journal of Applied Psychology, 66, 451-457.

Mandell, M.N. (1956). Supervisory characteristics and rating. Personnel, 32, 435-440.

March, J.C., & March, J.G. (1978). Performance sampling in social matches. Administrative Science Quarterly, 23, 434-453.

Mumford, M.D. (1983). Social comparison theory and the evaluation of peer evaluations: A review and some applied implications. Personnel Psychology, 36, 867-881.

Murphy, K.R., Garcia, M., Kerkar, S., Martin, C., & Balzer, W.K. (1982). Relationship between observational accuracy and accuracy in evaluating performance. Journal of Applied Psychology, 67, 320-325.

Nathan, B.R., & Lord, R.G. (1983). Cognitive categorization and dimensional schemata: A process approach to the study of halo in performance ratings. Journal of Applied Psychology, 68, 102-114.

Newtson, D.A. (1976). Foundations of attribution: The perception of ongoing behavior. In Harvey, et. al. (Eds.), New directions in attribution research, Vol. 1. Hillsdale, NJ: Erlbaum.

Norton, S.D., Gustafson, D.P., & Foster, C.E. (1977). Assessment for management potential: Scale design and development, training effects and rater/ratee sex effects. Academy of Management Journal, 20, 117-131.

Roadman, H.E. (1964). An industrial use of peer ratings. Journal of Applied Psychology, 48, 211-214.

Smith, P.C., & Kendall, L.M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 47, 149-155.

Taylor, S.E., & Crocker, J. (1981). Schematic bases of social information processing. In E. Higgins, C. Herman, & M. Zanna (Eds.), Social cognition: The Ontario symposium (Vol. 1). Hillsdale, NJ: Erlbaum.

Warmke, D.L., & Billings, R.S. (1979). Comparison of training methods for improving the psychometric quality of experimental and administrative performance ratings. Journal of Applied Psychology, 64, 124-131.

Wexley, K.N., & Klimoski, R. (1984). Performance appraisal: An update. In K.M. Rowland & G.R. Ferris (Eds.), Research in personnel and human resources management, Vol. 2. Greenwich, CT: JAI Press, Inc., 35-79.

Wherry, R.J., & Fryer, D.H. (1949). Buddy ratings: Popularity contest or leadership criteria? Personnel Psychology, 2, 147-159.

Whitla, D.K., & Tirrell, J.E. (1953). The validity of ratings of several levels of supervisors. Personnel Psychology, 6, 461-466.

Zedeck, S., & Baker, H.T. (1972). Nursing performance as measured by behavioral expectation scales: A multitrait-multirater analysis. Organizational Behavior and Human Performance, 7, 457-466.

Zedeck, S., & Kafry, D. (1977). Capturing rater policies for processing evaluation data. Organizational Behavior and Human Performance, 18, 269-294.

Subject Training Guide Outline

I. Explanation of Experimental Tasks

II. Exercise to Differentiate betwen Behavioral Incidents and Evaluations

III. Definition of Dimensions

A. Read

B. Example given

C. Any questions from the subject

Appendix A

Definitions of Videotape Dimensions

Delivery

The presenter's manner of speaking and the extent to which he or she uses audiovisula aids to clarify and emphasize important points of his of her presentation.

Depth of Knowledge

The presenter's mastery of the subject matter; this includes how well he or she knows the literature and thr research he or she reports.

Interpersonal Relations with Students

The presenter's rapport with and sensitivity to the audience and their questions.

Organization

The presenter's arrangement of the lecture material; the extent to which the presenter leads the class through a logical and orderly sequence of material.

Relevance

The presenter's choice of examples used in conveying information; examples which are important and meaningful to the audience.

Post-experimental Narrative Questions

1. How did you decide to press the button to record your observance of a behavioral incident?

2. How did you decide what rating (1 = very negative, 4 = very positive) to give each observed incident?

3. Do you believe that your role relationship to the presenter (peer or supervisor) influenced what you observed? If so, how? If not, why not?

4. Please rate the presenter on a 1 - 7 scale (1 = very negative, 7 = very positive) for each of the five dimensions on the sheet given you.

Appendix C

DEBRIEFING

The purpose of this research is to examine peer (i.e. undergraduate students) and supervisor (i.e. graduate teaching assistants) differences in the observation of performance-related behaviors. To do this, half of the subjects are peers of the student in the videotape and half are supervisors. All subjects view the same videotape of an undergraduate student giving an oral classroom presentation. By asking the subjects to stop the vedeotape whenever they observe a behavior indicative of a certain performance dimension, we hope to isolate observation from the later cognitive processes of encoding, storage, retrieval, and integration.

The results of this study may have important implications for ratings. If there are observational differences due to the rater's position relative to the student (rater as a peer or a supervisor), ratings may be obtained from raters with these different perspectives to provide more varied and complete information about the ratee's job performance.

Data for this study are being collected this semester (Spring 1985). Therefore, we have no results to report to you at this time. If you are interested in obtaining the results of this study, or have further questions about the research, please feel free to contact either of us listed below. We hope that your participation was both educational and interesting.

Mary W. Daley
William K. Balzer, Ph.D.

Appendix D

END

FILMED

12-85

DTIC